R-Laboratory 7

MANOJ KUMAR - 2048015

12/03/2021

### 1.Download the dataset BOSTON.csv

# Loading BOSTON dataset.  
  
dataset <- read.csv("boston.csv")  
head(dataset)

## TOWN TRACT LON LAT MEDV CRIM ZN INDUS CHAS NOX RM AGE  
## 1 Nahant 2011 -70.9550 42.2550 24.0 0.00632 18 2.31 0 0.538 6.575 65.2  
## 2 Swampscott 2021 -70.9500 42.2875 21.6 0.02731 0 7.07 0 0.469 6.421 78.9  
## 3 Swampscott 2022 -70.9360 42.2830 34.7 0.02729 0 7.07 0 0.469 7.185 61.1  
## 4 Marblehead 2031 -70.9280 42.2930 33.4 0.03237 0 2.18 0 0.458 6.998 45.8  
## 5 Marblehead 2032 -70.9220 42.2980 36.2 0.06905 0 2.18 0 0.458 7.147 54.2  
## 6 Marblehead 2033 -70.9165 42.3040 28.7 0.02985 0 2.18 0 0.458 6.430 58.7  
## DIS RAD TAX PTRATIO  
## 1 4.0900 1 296 15.3  
## 2 4.9671 2 242 17.8  
## 3 4.9671 2 242 17.8  
## 4 6.0622 3 222 18.7  
## 5 6.0622 3 222 18.7  
## 6 6.0622 3 222 18.7

- LON and LAT are the longitude and latitude of the center of the census tract.  
- MEDV is the median value of owner-occupied homes, measured in thousands of dollars.  
- CRIM is the per capita crime rate.  
- ZN is related to how much of the land is zoned for large residential properties.  
- INDUS is the proportion of the area used for industry.  
- CHAS is 1 if a census tract is next to the Charles River else 0   
- NOX is the concentration of nitrous oxides in the air, a measure of air pollution.  
- RM is the average number of rooms per dwelling.  
- AGE is the proportion of owner-occupied units built before 1940.  
- DIS is a measure of how far the tract is from centres of employment in Boston.  
- RAD is a measure of closeness to important highways.  
- TAX is the property tax per $10,000 of value.  
- PTRATIO is the pupil to teacher ratio by town

str(dataset)

## 'data.frame': 506 obs. of 16 variables:  
## $ TOWN : chr "Nahant" "Swampscott" "Swampscott" "Marblehead" ...  
## $ TRACT : int 2011 2021 2022 2031 2032 2033 2041 2042 2043 2044 ...  
## $ LON : num -71 -71 -70.9 -70.9 -70.9 ...  
## $ LAT : num 42.3 42.3 42.3 42.3 42.3 ...  
## $ MEDV : num 24 21.6 34.7 33.4 36.2 28.7 22.9 22.1 16.5 18.9 ...  
## $ CRIM : num 0.00632 0.02731 0.02729 0.03237 0.06905 ...  
## $ ZN : num 18 0 0 0 0 0 12.5 12.5 12.5 12.5 ...  
## $ INDUS : num 2.31 7.07 7.07 2.18 2.18 2.18 7.87 7.87 7.87 7.87 ...  
## $ CHAS : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ NOX : num 0.538 0.469 0.469 0.458 0.458 0.458 0.524 0.524 0.524 0.524 ...  
## $ RM : num 6.58 6.42 7.18 7 7.15 ...  
## $ AGE : num 65.2 78.9 61.1 45.8 54.2 58.7 66.6 96.1 100 85.9 ...  
## $ DIS : num 4.09 4.97 4.97 6.06 6.06 ...  
## $ RAD : int 1 2 2 3 3 3 5 5 5 5 ...  
## $ TAX : int 296 242 242 222 222 222 311 311 311 311 ...  
## $ PTRATIO: num 15.3 17.8 17.8 18.7 18.7 18.7 15.2 15.2 15.2 15.2 ...

summary(dataset)

## TOWN TRACT LON LAT   
## Length:506 Min. : 1 Min. :-71.29 Min. :42.03   
## Class :character 1st Qu.:1303 1st Qu.:-71.09 1st Qu.:42.18   
## Mode :character Median :3394 Median :-71.05 Median :42.22   
## Mean :2700 Mean :-71.06 Mean :42.22   
## 3rd Qu.:3740 3rd Qu.:-71.02 3rd Qu.:42.25   
## Max. :5082 Max. :-70.81 Max. :42.38   
## MEDV CRIM ZN INDUS   
## Min. : 5.00 Min. : 0.00632 Min. : 0.00 Min. : 0.46   
## 1st Qu.:17.02 1st Qu.: 0.08205 1st Qu.: 0.00 1st Qu.: 5.19   
## Median :21.20 Median : 0.25651 Median : 0.00 Median : 9.69   
## Mean :22.53 Mean : 3.61352 Mean : 11.36 Mean :11.14   
## 3rd Qu.:25.00 3rd Qu.: 3.67708 3rd Qu.: 12.50 3rd Qu.:18.10   
## Max. :50.00 Max. :88.97620 Max. :100.00 Max. :27.74   
## CHAS NOX RM AGE   
## Min. :0.00000 Min. :0.3850 Min. :3.561 Min. : 2.90   
## 1st Qu.:0.00000 1st Qu.:0.4490 1st Qu.:5.886 1st Qu.: 45.02   
## Median :0.00000 Median :0.5380 Median :6.208 Median : 77.50   
## Mean :0.06917 Mean :0.5547 Mean :6.285 Mean : 68.57   
## 3rd Qu.:0.00000 3rd Qu.:0.6240 3rd Qu.:6.623 3rd Qu.: 94.08   
## Max. :1.00000 Max. :0.8710 Max. :8.780 Max. :100.00   
## DIS RAD TAX PTRATIO   
## Min. : 1.130 Min. : 1.000 Min. :187.0 Min. :12.60   
## 1st Qu.: 2.100 1st Qu.: 4.000 1st Qu.:279.0 1st Qu.:17.40   
## Median : 3.207 Median : 5.000 Median :330.0 Median :19.05   
## Mean : 3.795 Mean : 9.549 Mean :408.2 Mean :18.46   
## 3rd Qu.: 5.188 3rd Qu.:24.000 3rd Qu.:666.0 3rd Qu.:20.20   
## Max. :12.127 Max. :24.000 Max. :711.0 Max. :22.00

# Checking for missing values.  
colSums(is.na(dataset))

## TOWN TRACT LON LAT MEDV CRIM ZN INDUS CHAS NOX   
## 0 0 0 0 0 0 0 0 0 0   
## RM AGE DIS RAD TAX PTRATIO   
## 0 0 0 0 0 0

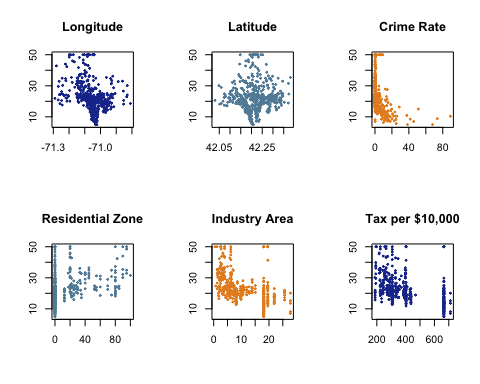
colSums(dataset=='')

## TOWN TRACT LON LAT MEDV CRIM ZN INDUS CHAS NOX   
## 0 0 0 0 0 0 0 0 0 0   
## RM AGE DIS RAD TAX PTRATIO   
## 0 0 0 0 0 0

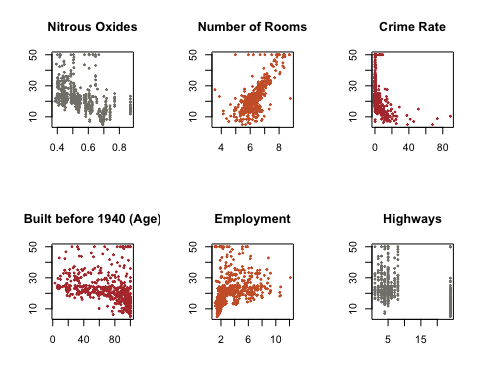
*Insight*

### 2.MEDV is the output /target variable i.e price of the house to be predicted.

par(mfrow=c(2,3))  
  
plot(dataset$LON, dataset$MEDV, main = "Longitude", xlab="", ylab="", cex.lab="1.5", col="#1e3799", pch=20, cex= 0.5 )   
plot(dataset$LAT, dataset$MEDV, main = "Latitude", xlab="", ylab="", cex.lab="1.5", col="#628ca6", pch=20, cex= 0.5)  
plot(dataset$CRIM, dataset$MEDV, main = "Crime Rate", xlab="", ylab="", cex.lab="1.5", col="#e58e26", pch=20, cex= 0.5)  
plot(dataset$ZN, dataset$MEDV, main = "Residential Zone", xlab="", ylab="", cex.lab="1.5", col="#628ca6", pch=20, cex= 0.5)  
plot(dataset$INDUS, dataset$MEDV, main = "Industry Area", xlab="", ylab="", cex.lab="1.5", col="#e58e26", pch=20, cex= 0.5)  
plot(dataset$TAX, dataset$MEDV, main = "Tax per $10,000", xlab="", ylab="", cex.lab="1.5", col="#1e3799", pch=20, cex= 0.5)



par(mfrow=c(2,3))  
  
plot(dataset$NOX, dataset$MEDV, main = "Nitrous Oxides", xlab="", ylab="", cex.lab="1.5", col="#84817a", pch=20, cex= 0.5 )   
plot(dataset$RM, dataset$MEDV, main = "Number of Rooms", xlab="", ylab="", cex.lab="1.5", col="#cd6133", pch=20, cex= 0.5 )   
plot(dataset$CRIM, dataset$MEDV, main = "Crime Rate", xlab="", ylab="", cex.lab="1.5", col="#b33939", pch=20, cex= 0.5 )   
plot(dataset$AGE, dataset$MEDV, main = "Built before 1940 (Age)", xlab="", ylab="", cex.lab="1.5", col="#b33939", pch=20, cex= 0.5 )   
plot(dataset$DIS, dataset$MEDV, main = "Employment", xlab="", ylab="", cex.lab="1.5", col="#cd6133", pch=20, cex= 0.5 )   
plot(dataset$RAD, dataset$MEDV, main = "Highways", xlab="", ylab="", cex.lab="1.5", col="#84817a", pch=20, cex= 0.5 )

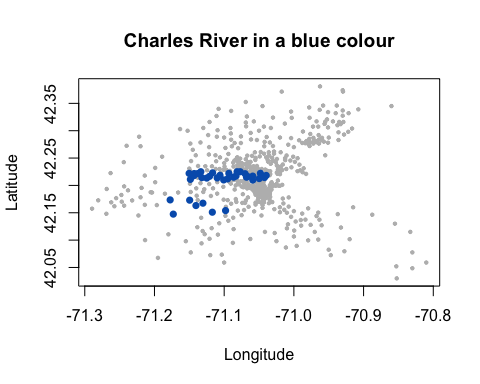


*Insight*

### 3. Using the plot commands, plot the latitude and longitude of each of our census tracts.

### 4. Show all the points that lie along the Charles River in a blue colour.

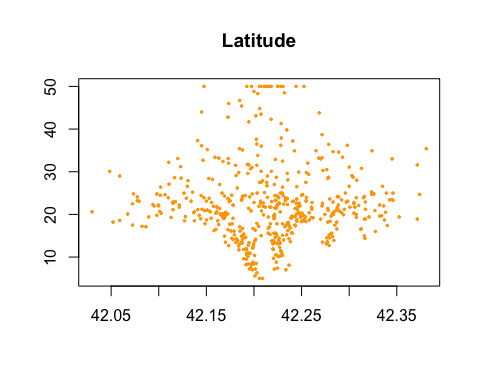
# https://colorhunt.co/  
  
plot(dataset$LON, dataset$LAT, main = "Charles River in a blue colour", xlab="Longitude", ylab="Latitude", col="#bbbbbb", pch=20, cex= 0.7)  
points(dataset$LON[dataset$CHAS==1], dataset$LAT[dataset$CHAS==1], col="#005fba", pch=16)



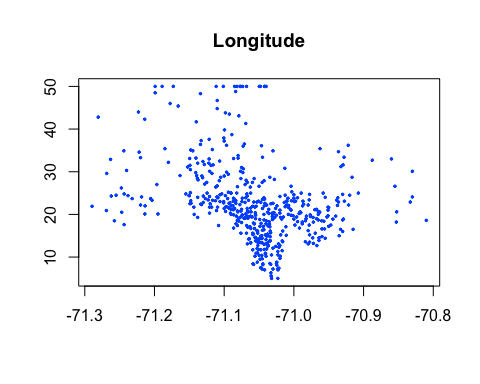
*Insight*

### 5.Apply Linear Regression by plotting the relationship between latitude and house prices and the longitude and the house prices.

plot(dataset$LAT, dataset$MEDV, main = "Latitude", xlab="", ylab="", cex.lab="1.5", col="#ffa500", pch=20, cex= 0.5)



plot(dataset$LON, dataset$MEDV, main = "Longitude", xlab="", ylab="", cex.lab="1.5", col="#005aff", pch=20, cex= 0.5)



#linear model  
linear\_model = lm(MEDV ~ LAT + LON, data=dataset)  
summary(linear\_model)

##   
## Call:  
## lm(formula = MEDV ~ LAT + LON, data = dataset)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -16.460 -5.590 -1.299 3.695 28.129   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -3178.472 484.937 -6.554 1.39e-10 \*\*\*  
## LAT 8.046 6.327 1.272 0.204   
## LON -40.268 5.184 -7.768 4.50e-14 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 8.693 on 503 degrees of freedom  
## Multiple R-squared: 0.1072, Adjusted R-squared: 0.1036   
## F-statistic: 30.19 on 2 and 503 DF, p-value: 4.159e-13

*Insight*

### 6.Apply Regression Tree to the problem and draw conclusions from it.

*Insight*